White Paper Report

Report ID: 100987

Application Number: HD5113810

Project Director: James Ginther (ginthej@slu.edu)

Institution: St. Louis University

Reporting Period: 9/1/2010-11/30/2011

Report Due: 2/29/2012

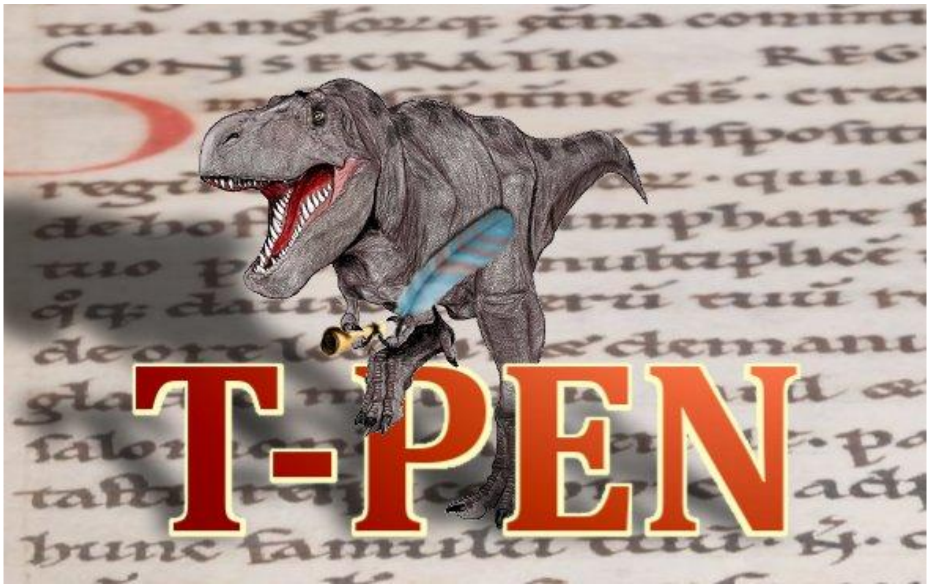
Date Submitted: 3/12/2012

White Paper
For Grant Number HD5113810

The T-PEN Tool: Sustainability and Quality Control in Encoding Handwritten Texts

James R. Ginther (PI)
Center for Digital Theology
Saint Louis University

Abigail Firey (Co-PI)
Department of History
University of Kentucky



http://www.t-pen.org

Introduction

As digital humanities come to include more and more large projects that incorporate transcriptions of handwritten texts (e.g., the Newton Project, the Petrarch Project, the Melville Project, the St. Patrick's Confessio Project, the Dickinson Project, the Codex Sinaiticus Project, etc.) and as many such projects are testing the possibilities of effective crowd-sourcing as a means to generate the transcriptions, two challenges are especially evident. First, there are multiple issues that can be collected under the general rubric of "sustainability": in brief, will the data and the project be properly curated for longevity? Second, especially in projects designed for scholarly research, where the quality of the data must be of the highest order, how is it possible to design and maintain good systems for generating good data and verifying that it is, in fact, good? "Good data" for the humanist is, more often than not, an accurate representation of the text. When the text(s) to be represented exist in handwritten exemplaria, standards of accuracy devolve to the modern transcriber, who becomes responsible for recording every incorrect as well as correct letter, every instance of addition, deletion, revision, and glossing in the original document, and, in some instances, peculiarities of the physical medium of the document itself, if the text has been affected (an example familiar to medievalists is that often manuscript pages have been trimmed for rebinding, and text once present is now excised or, quite literally, truncated).

With the generous support of an NEH DH Start-up Grant (Level 2) and aconcurrent one from the Mellon Foundation, we have been able to incorporate our recognition of the gravity of these two challenges – sustainability and quality control-- into a transcription tool, T-PEN (Transcription for Paleographical and Editorial Notation). We developed T-PEN initially for the transcription of digitized images of medieval manuscripts delivered publically by manuscript repositories, but T-PEN can be used with images of handwritten documents of any era. In the following pages, we detail the features of the tool that allow digital humanists to respond to the challenges of sustainability and quality control more effectively.

1. The Hand and Eye

As medievalists, we were trained to recognize the types of errors that transcribers (who are, after all, akin to medieval scribes) are most prone to make: haplography (omission of content between similar or identical words; "saut du même au meme"), dittography (repetition of letters or syllables), duplication or omission (of letters, words, or lines), often caused by homoearcton and homoeoteleuton (similar beginnings and endings of words), and transpositions. Many of these errors are caused by the simple mechanics of the transcriber placing his gaze at a slightly different point in the text when he shifts it back from his own transcription to his exemplar. To reduce the likelihood of such errors, we took advantage of digital technology to place both the transcription and the exemplar so that visual movement between the two is as minimal as possible, and the field of vision is as controlled as possible. This we accomplished with a simple but novel visualization of the lines of script in the exemplar, which we integrated with interactive transcription spaces. To build the tool, we developed an algorithm for "parsing" the lines of script in an image, and a data model that connected the image delivery of manuscript repositories with the actions of transcribers.

2. Project Background: The Line Parsing Algorithm

For the visualization of the lines of script in the image to be transcribed, we drew upon a prototype that had been created for the Electronic Norman Anonymous Project, a digital edition project funded by the Andrew Mellon Foundation that was completed in July 2010 at the Center for Digital Theology (CDT) of Saint Louis University (Ginther and O'Sullivan 2010). For that project, the initial aim was to develop a tool that permitted users to critique fully the edition of an early twelfth-century Latin text, both by inserting marginal comments in the edition, and also using the transcription tool to provide access to images of the sometimes problematic letter forms and abbreviations of the single manuscript witness. This prototype thus already offered means for parsing each individual line of a manuscript page, so that it could then be displayed alongside the transcription of that line. The accuracy of the parsing was extensively tested (including a large crowd-sourcing review) during 2010, over increasing numbers of digital images of medieval manuscripts, beginning with 20 Middle English manuscripts, then images in the e-codices project (www.e-codices.unifr.ch), and then, with the support of a small grant from the Saint Louis University President's Research Fund, 506 manuscripts that comprise the *Parker on the Web* collection (parkerweb.stanford.edu), as well as a sampling of digitized images from the Vatican Film Library, Saint Louis University. Overall the sample space was comprised of 196,883 individual page images that contained text written in Latin, medieval German and French, Old and Middle English and New Testament Greek. Accuracy rates varied from 90% to 73%. We were confident, however, that the line parsing algorithm could become the foundational feature of a more advanced transcription tool. In the summer of 2011, after further work on the parsing algorithm, the accuracy rate rose to 80%.

In very basic terms, T-PEN's line parsing algorithm works with the assumption of "dark and light" contrast between where handwriting is extant and where it is absent (Likforman-Sulem 2007). It identifies the coordinates for each line of a handwritten page in the following steps (see Figures 1 and 2):

(1) The digital image of the manuscript page (which is normally in full color) is rendered in contrasted black and white. The image is also resized to 1000px in height but maintains the aspect ratio.

(2) The number of columns are first determined by locating the “margins” of the writing space, that is where there are few or no non-white objects. A heavily glossed manuscript may return a greater number of columns than may actually exist, but for the most part column identification returns an accurate set of coordinates for where columns begin and end both horizontally and vertically. The column area is displayed as a shaded area in T-PEN’s User interface.

(3) The algorithm smears the image horizontally to reinforce the distinction between areas with writing and those without; it then passes through the rows of pixels that comprise a column, determining where a top of a line begins by testing for a threshold of black pixels. A series of increasing amount of black pixels (e.g. row_{n} = 15% black pixels, row_{n+1} = 20% black, row_{n+2} = 40% black, etc.) is read as the top of a line. The algorithm then tests for a diminishing threshold to determine where the bottom of the line is. This process is iterated until the algorithm reaches the coordinates that map the bottom of the column. The algorithm then completes this process for each additional column on the page.

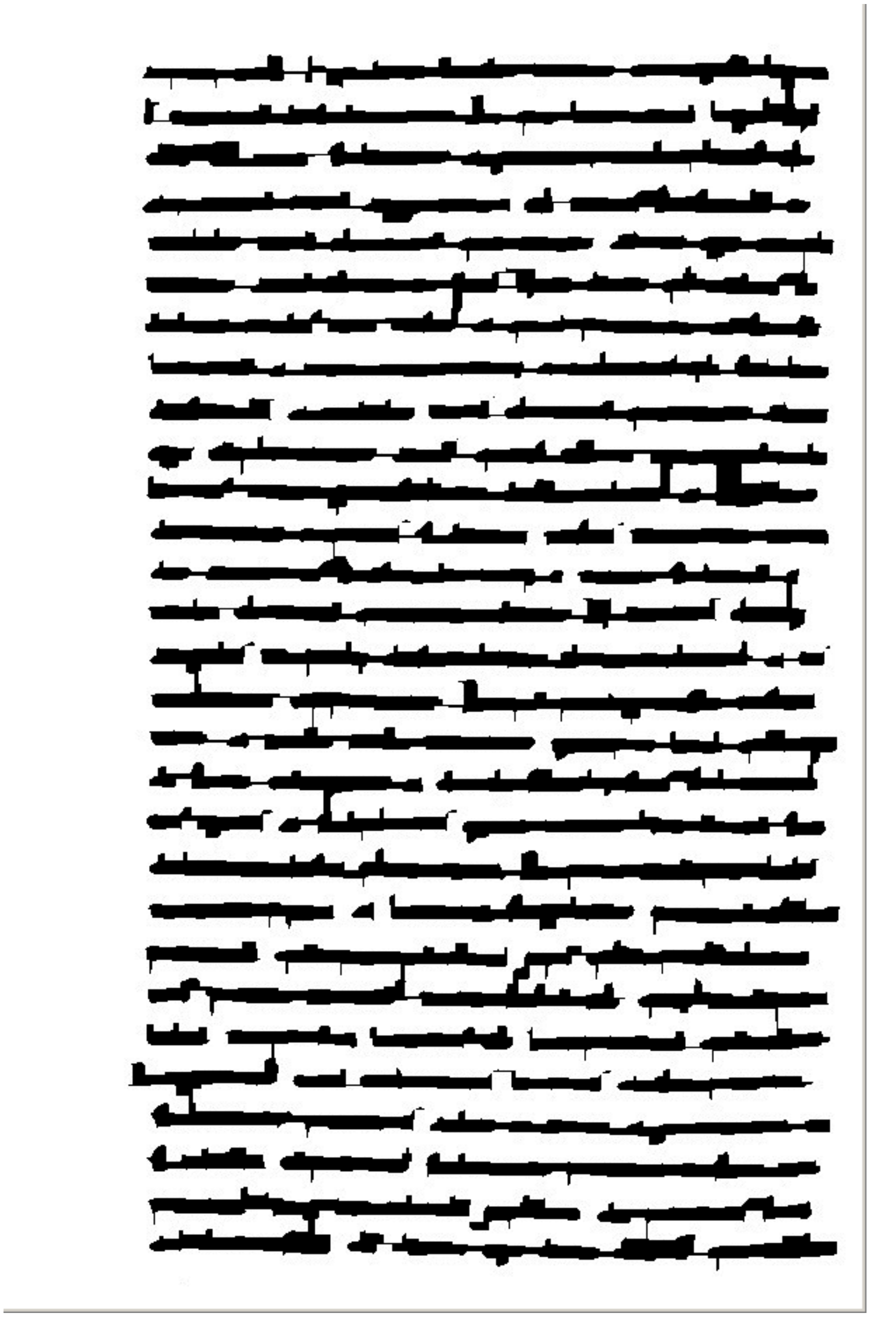


Figure 1 A smeared version of the binary threshold of the manuscript image (Cambridge, Corpus Christi College MS 415, p. 22)

(4) The resulting coordinates are stored in a matrix that can be passed to any GUI function that will display the image of the page.

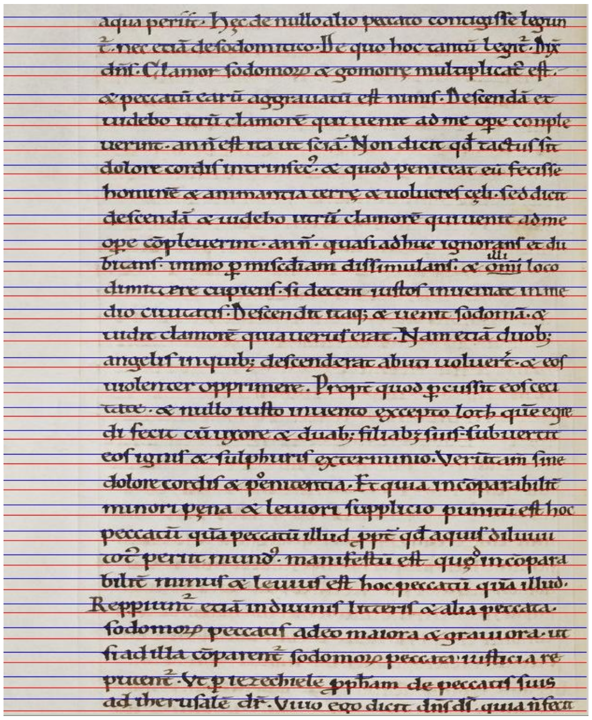


Figure 2 A Parsed Manuscript Page (Cambridge, Corpus Christi College MS 415, p. 22: the blue line indicates the detected top of a written line; the red line indicates the detected bottom of a written line.

Even a well developed algorithm will, given the wide range of highly idiosyncratic forms of manuscript pages, sometimes deliver results that may not be satisfactory, and some transcribers have particular preferences for the visual representation of script lines. T-PEN therefore offers transcribers the option of manually revising the parsing of both columns and lines in its display. When any manual corrections are made to the line parsing in a project, those changes in the parsing coordinates are part of the individual project's transcription record. In this way, there can be multiple parsings, creatd by multiple transcribers, for any page image. Manual alteration of line parsing never affects the default parsing that is associated with each public display of the image.

3. Project Background: T-PEN's Data Model

Complementing T-PEN's capacity to display visually clarified lines of script for transcription is the data model, which is composed of four objects: image, manuscript, project and transcription. These objects and their fields are as follows:

The **Image Object** contains the URL that points to the originating repository for each page image, and a referent to the manuscript object from which the individual page image comes. When these images are parsed for the first time, that matrix of coordinates also becomes part of the image object and will be the public, default parsing for each manuscript page.

The **Manuscript Object** contains the list of the relevant image objects that comprise the manuscript. It also has a field for the shelf-mark (the unique names used in cataloguing manuscripts) by which this manuscript is identified at the originating repository. T-PEN has adopted the "SharedCanvas" model for consuming manuscript manifests from digital repositories. SharedCanvas is a RDF Linked Data model for establishing the content of a manuscript, and connecting it to annotations and transcriptions that derive from those digital images (Sanderson et al 2011). SharedCanvas utilizes the Open Annotation Consortium (OAC) model for attaching images to canvases. T-PEN also uses OAC to share transcriptions with other OAC-aware applications (Halshofer et al 2011). While we have adopted SharedCanvas, T-PEN is able to consume details about manuscripts from repositories that use other metadata configurations with some customization of the existing codebase to meet the repository's exposed configuration.

The Manuscript Object can also have a second component. In cases of "restricted" manuscripts (that is manuscripts which are designated only for a specific set of users), the name of the T-PEN user who manages permissions and the names of other users who have been granted access are also part of this object. This means that T-PEN maintains a distinction between granting transcribers access to a digitized manuscript and granting transcribers access to a transcription.

The **Project Object** comprises the data about each project created in T-PEN. In a project, there are two required fields: a T-PEN user and some number of image objects. The T-PEN user may be a single scholar, or may be the leader of a collaboration with multiple transcribers, whom the leader can add or dismiss at any time. The images typically comprise a single manuscript with the page images in the order originally specified by the repository. Within T-PEN, however, it is possible both to add images from multiple manuscripts and reorder the images, to allow the creation of a virtual manuscript. This feature is especially important when the scholar knows or hypothesizes that a manuscript may have become "dismembered" (the technical term) in the course of rebinding or as a result of historical trauma, and wishes to reconstruct the original manuscript virtually. The Project Object also contains additional details or options for project configuration to be discussed below, such as association of a RelaxNG schema, connection to an external digital project to which transcriptions will be exported from T-PEN, customized buttons for inserting XML markup in transcriptions, and association of a Dublin Core or TEI header .

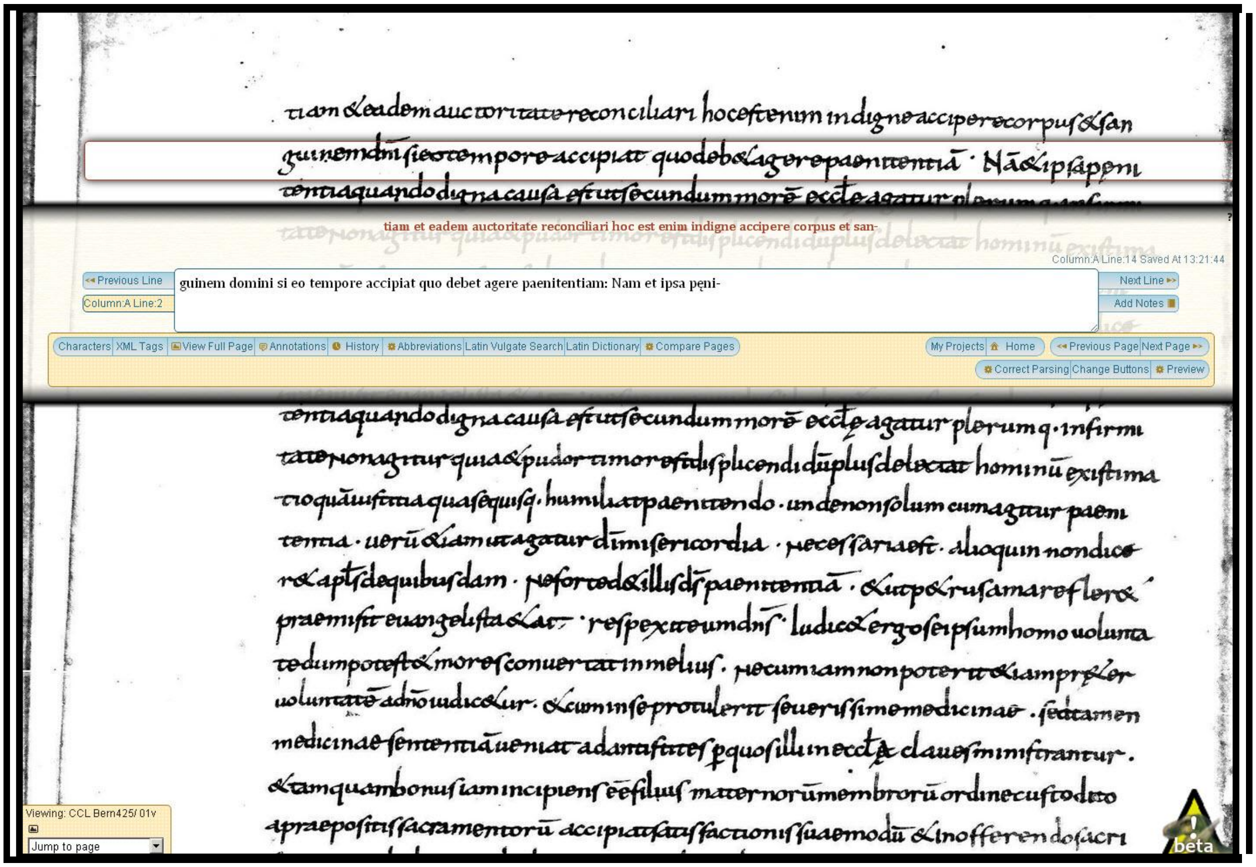
Finally, the **Transcription Object** is the fundamental component of T-PEN's data model. Each instantiation contains a referent to the image object, the coordinates for a single line, the character data of the transcription (along with any XML markup transcribers have inserted), the character data of any notes attached to the line, and a referent to the project object to which it is connected.

4. User experience

With the line parsing algorithm and the data model conceptually in place so that we could reliably supply the images of manuscript pages for transcription, we developed a user interface that could enhance significantly the accuracy of transcriptions and also support efficient proofreading. In addition to designing an interface that users would find both intuitive and flexible enough to accommodate the differing ocular strategies – focusing, magnifying, surveying, contextualizing, comparing various aspects of the handwritten document -- of individual transcribers, we incorporated additional resources that scholars use to check and improve their perception of the letter forms: dictionaries for Latin and medieval vernacular languages, a Latin Vulgate Bible, and a dictionary of Latin manuscript abbreviations.

(a) The Transcription Environment

T-PEN's main user interface (UI) reflects our commitment to creating software that seems intuitive in its use, even to scholars accustomed to working in a print—or manuscript! – environment, while taking advantage of digital technology to improve the activity of transcribing handwritten documents. As noted above, our goal was to reduce the visual distance between the line of script to be transcribed and the transcription in progress. The T-PEN UI allows the user to see simultaneously the page image and the transcription space, which is presented as a dynamic overlay that the transcriber moves over the image, line by line. The full screen image of the page offers a demarcation of the line to be transcribed by a bounding box drawn in red, which works well on most tonalities of parchment, paper, and inks. The red box can be dismissed by holding down the CTRL key, should it obscure features the transcriber wishes to view more closely.. Beneath the bounded line hangs a "transcription box" in which the transcriber enters the transcription of the line immediately above; thus the field of view is set for close reference of the manuscript line and the transcription line. A second box for notes can also be opened immediately below the transcription box, so that the scholar can make a running record of comments and observations that relate to the line immediately in view. Users can then navigate to the next or previous line; upon moving to the next line, the transcription of the previous line is displayed in a muted red above the transcription box, and any notes the scholar added are displayed in grey. This allows the transcriber to read the text as it accrues, an essential aspect of good transcription. As the transcriber progresses (or moves back through the page) the page image is adjusted accordingly. In this way, almost the entire image is always in view, as it is important to see other lines on the page, both in order to retain one's sense of one's location in the text and to confirm readings and the practices of the original scribe.



tiam et eadem auctoritate reconciliari hoc est enim indigne accipere corpus et san-

guinem domini si eo tempore accipiat quo debet agere paenitentiam: Nam et ipsa peni-

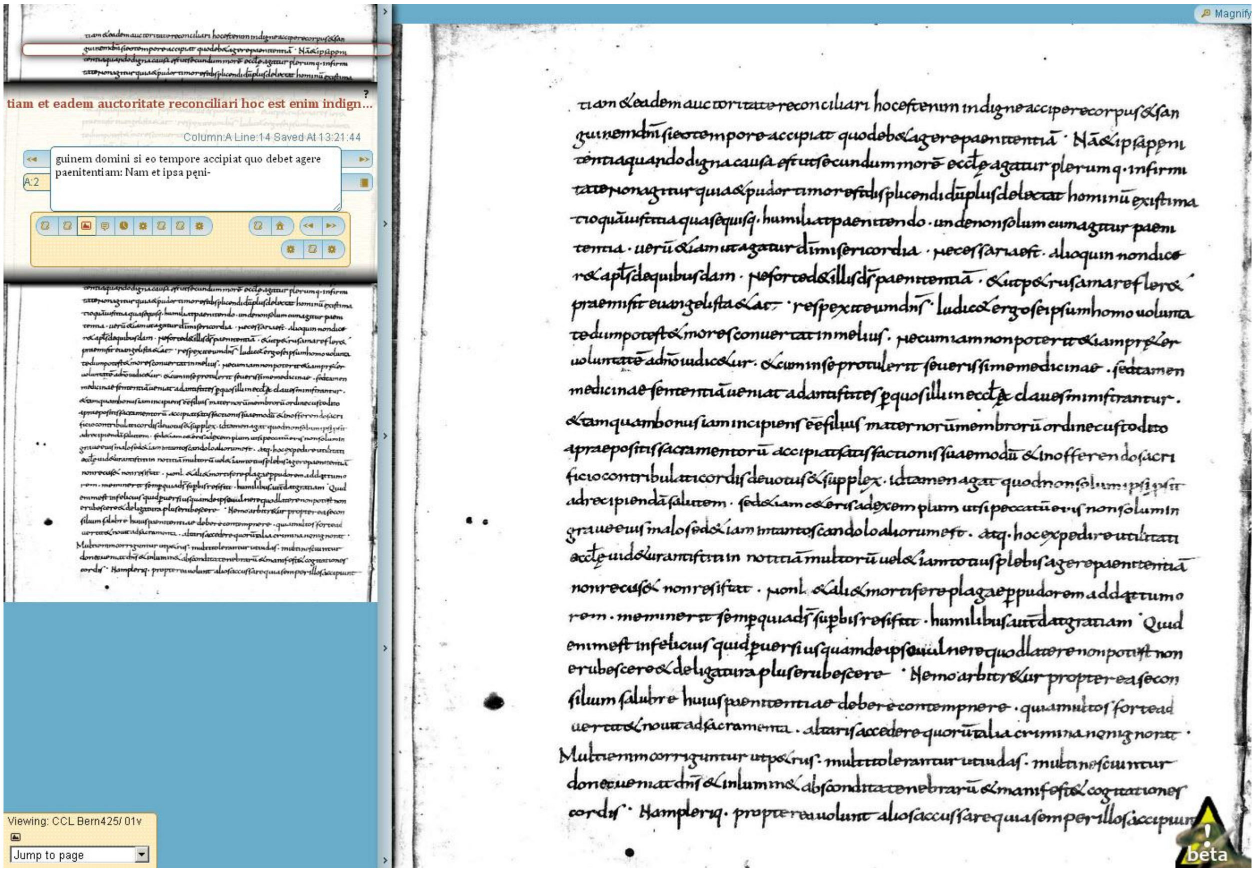
Column:A Line:14 Saved At 13:21:44

Column:A Line:2

Viewing: CCL Bern425/ 01v

Figure 3 T-PEN's Transcription UI

Often the page image is made up of two or three (occasionally even four) columns, or is so large that it is not possible to present the entire page image this mode. T-PEN allows the transcriber to open a split-screen copy of the whole page. This serves not only the purpose of viewing the whole page, but also provides short-cut navigation. In the normal transcription box view, users must navigate one line at a time. If a transcriber wishes to jump 10 lines ahead, she must move through the intervening 9 lines to get there. In “View Full Page” mode, she can mouse over the line she wishes to view and click on it. The transcription box is immediately reoriented to that line, and when the full page mode is dismissed the box remains at the new line.



tiam et eadem auctoritate reconciliari hoc est enim indign...

Column:A Line:14 Saved At 13:21:44

guinem domini si eo tempore accipiat quo debet agere paenitentiam: Nam et ipsa pęni-

Figure 4 "View Full Page" mode in T-PEN

The T-PEN UI underwent extensive testing and development. Perhaps our greatest discovery was that we needed to respond to the subtle but important habits of transcribers. We needed to respect their resistance to conforming to the tool; rather, we needed to make the tool conform to their desires. Experienced transcribers, for example, were unwilling to revert to the strict line-by-line practices of their early days in paleography classes; they preferred to transcribe by sentence or sense units. Giving users the power to set their own line-parsing patterns means that such transcribers can bound two or three lines, rather than one, as the unit within the red bounding box associated with the transcription box. We also adapted the UI to accommodate the habits of scholars used to working through transcriptions quickly, who were annoyed by the time and effort needed to mouse to the "next line" tab: T-PEN allows such transcribers to use the typist's standard "hit return/enter" to advance a line. We also learned how much transcribers roam over the page, or over several pages, as they make decisions about how to transcribe problematic scripts or features on the page, and with the funding from the Mellon Foundation we extended the "view full page" feature to support "compare pages" on a split screen. We also put "next page" and "previous page" tabs on the transcription box, adjacent to the "next line" and "previous line" tabs. Even the bounding box evolved into a form that has visual grace and dynamism: users love the little "bounce" it gives when it resituates itself on a new line, because it gives a sense of smooth progression to a new location. We also repeatedly increased the speed with which images and features are delivered to the transcriber, in an effort to ensure that the T-PEN environment can match the traditional "books and papers on a table" environment for transcription in terms of efficiency.

Knowing that T-PEN would be invaluable as a tool for proofreading as well as for generating accurate transcriptions, with the support of the Mellon Foundation, we also developed the capacity for scholars to upload existing transcriptions (after putting them into plain text or xml format) to a T-PEN project so that they could, by using the T-PEN "line-breaking" feature, re-associate lines of transcription with lines of script in the manuscript image. In other words, transcribers can populate the transcription boxes with data not generated with T-PEN, and then use T-PEN for verifying the accuracy of that existing transcription, correcting it easily, annotating it in the T-PEN "notes" boxes, and inserting xml markup with T-PEN's "auto-encoding" feature (discussed below). The corrected transcription can then be exported as any T-PEN-generated transcription would be, and the images have a transcription associated with them for collaborative study.

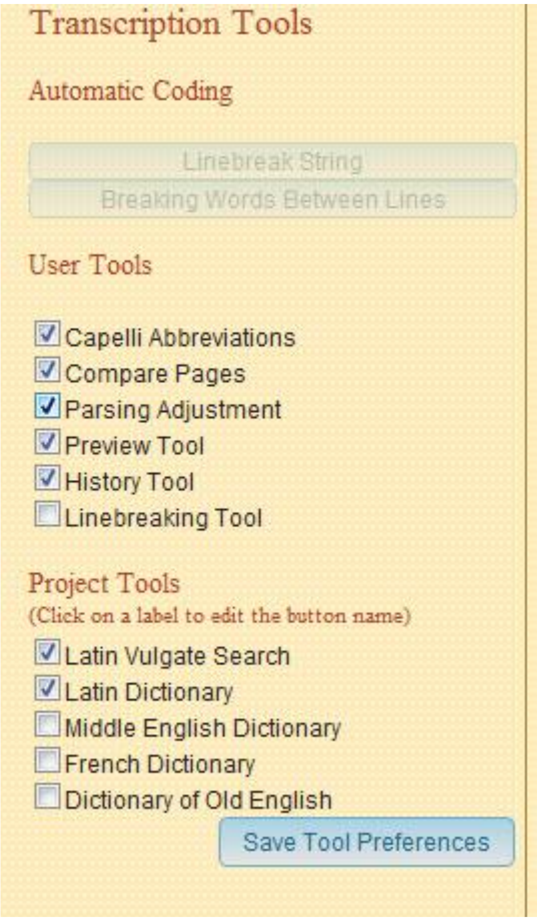
(b) The collaborative environment

As we developed T-PEN, we were always mindful that it would likely be used both by individuals preparing studies or editions on their own, possibly with a print publication as the ultimate objective, and scholars working on digital projects that would likely be larger, collaborative endeavours. In terms of quality control, collaboration is perhaps a double-edged sword. Collaborators can check each other's work, but if they are not engaged in mutual verification of the data, one person's errors may enter a large data set unnoticed. As the number of collaborators increases, so do the potential benefits and hazards. As we considered the implications of collaboration for quality control, we drew upon the experience of the Carolingian Canon Law (CCL) project, T-PEN's partner in development. The CCL is undertaking a task too large to be accomplished by a single investigator: it is publishing online transcriptions of several thousand manuscripts of early medieval canon law, to provide access to texts that remain largely unpublished. Because it is hoped that the CCL will help to break the impasse in the production of critical editions of these materials, which have largely defied the application of traditional editorial methods, it is essential that CCL transcriptions have to-the-letter accuracy, so that editors can use them with confidence. The corpus is so large, however, that it will require many contributors to build the database of transcriptions. The CCL therefore invites any and all interested scholars and students with registered accounts on the CCL to contribute either short or long transcriptions of canon law texts, as they find them and as they feel inspired to do so (Firey 2010). The first question usually heard about this method is, "but how will you ensure that the transcriptions are of sufficient quality?"

T-PEN provides an environment that both encourages collaborative verification of the accuracy of transcriptions and also makes it easy to accomplish that verification. As a web-based application, T-PEN is available to scholars and students all over the world, who may or may not be known to the immediate CCL team, but can now participate in contributing to the CCL's growth. Projects such as the CCL can now ask that contributors prepare their transcriptions in T-PEN, where CCL team members can be invited to a transcription project to proofread and correct any errors. When the CCL publishes transcriptions, it provides a public statement about the reliability of the transcription: it declares the level of the transcriber's experience, the quality of the manuscript and images, and whether the transcription has been proofread. The CCL can thus publish new data quickly, without waiting for the full, time-consuming process of verification to be completed. What we find, however, now that we are using T-PEN, is that we can complete proofreading and correction rapidly. The surprised and unanimous response has been that "proofreading with T-PEN is a dream!"

5. Transcription Tools

The final feature of quality control to be noted is the variety of tools that assist the transcriber in her work. Many of these tools reflect T-PEN's initial development for projects based upon medieval manuscripts written in Latin. A transcriber begins by selecting which scholarly reference works she would like to have available in the Transcription UI. This selection is made on the Options tab on T-PEN's project management page.



Transcription Tools

Automatic Coding

User Tools

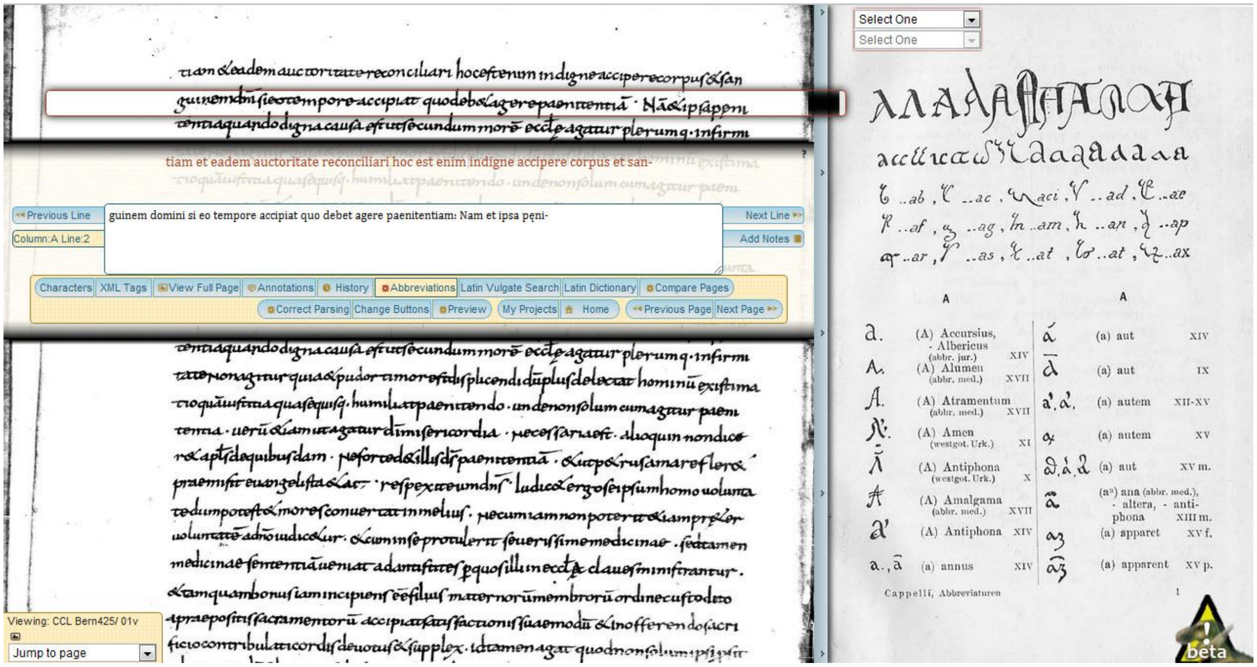
☑ Capelli Abbreviations
☑ Compare Pages
☑ Parsing Adjustment
☑ Preview Tool
☑ History Tool
□ Linebreaking Tool

Project Tools
(Click on a label to edit the button name)

☑ Latin Vulgate Search
☑ Latin Dictionary
□ Middle English Dictionary
□ French Dictionary
□ Dictionary of Old English

Figure 5 Tool Selection in T-PEN's project management

The chosen resources then appear on the bottom of the Transcription UI as labeled buttons. When a resource is selected by clicking a button, the browser screen is split, and the reference work is displayed on the right and the page image is displayed on the left. This permits the user to continue to view the transcription in process while using the associated scholarly reference work.



tiam et eadem auctoritate reconciliari hoc est enim indigne accipere corpus et san-

Column:A Line:2

guinem domini si eo tempore accipiat quo debet agere paenitentiam: Nam et ipsa pęni-

Viewing: CCL Bern425/ 01v

Figure 6 Displaying the Latin Abbreviations tool

As one can see in figure 5, T-PEN currently provides access to a Latin Dictionary (a local version of the dictionary from the Perseus project), Capelli's *Abbreviationes*, a searchable Latin Vulgate Bible (Clementine edition), and dictionaries for Medieval French, Old and Middle English.

The primary reason that T-PEN's development pursued these transcription tools was that they increase the quality of transcription. Scholars who may be new to Latin manuscripts can sometimes become overwhelmed by the system of abbreviations that comprises thousands of characters. Having Capelli's dictionary of abbreviations at their fingertips can go a long way to make even their first attempts at transcription a successful endeavor. Being able to search the text of the Latin Bible, perhaps the only true ubiquitous text in the Middle Ages, can also facilitate accurate transcription. The same is true for having access to language dictionaries. Since these tools are accessible through an i-frame configuration, it will be very easy to include additional tools as different types of repositories become available.

6. Re-integration of transcribing and encoding

At the most general level, the question of quality control in digital projects is related to conformity to accepted standards. Digital humanities has matured to the point of having sets of internationally recognized protocols, designed to reduce the proliferation of idiosyncratic datasets that can only be used in one project, or one environment, or by one investigator. For textual scholars, perhaps the most noteworthy set of protocols is the structured approach developed by the Text Encoding Initiative (TEI) for recording the textual and physical features in handwritten materials. Particularly in its most recent version (P5), TEI has formulated an XML encoding method that can assist transcribers in recording variant readings, corrections, ellipses, marginal glosses, etc., in a standardized manner (Burghart and Rehbein 2012; Driscoll 2006).

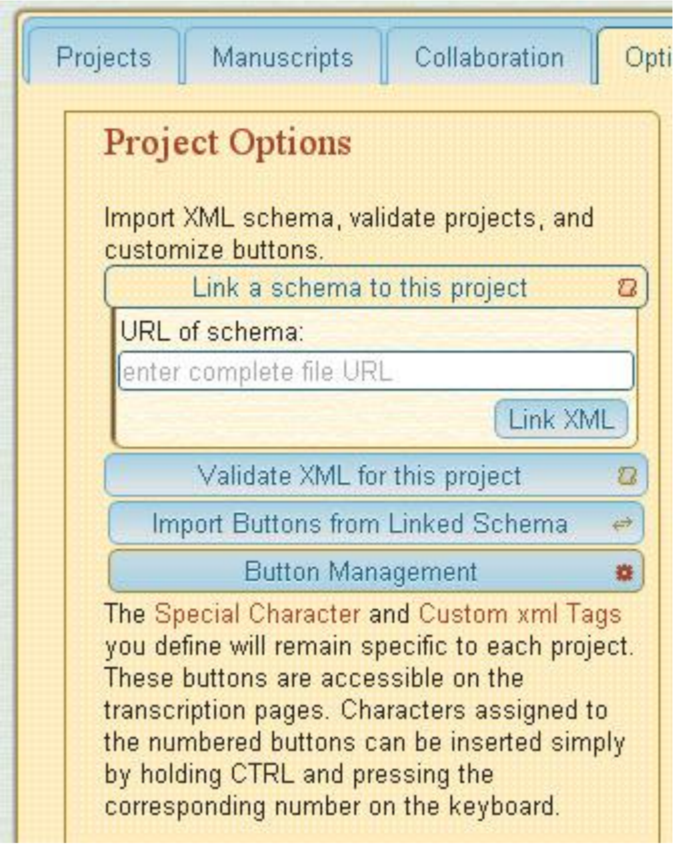
The rapid increase in digital transcription and editing, however, has yielded an unintended result, namely a separation of the processes of transcription and encoding. Digital preparation of a text becomes a two-step process, in which the editor creates an electronic transcription in one application and then transfers it to another in order to insert the XML encoding. This is a logical step as most XML editing applications do not provide a workspace that facilitates transcription work. The problem is that there can be a loss of accuracy (especially with regards to the textual and codicological attributes mentioned above) between the transcription created using the document *in situ* or a facsimile copy and the encoding process. There is even more danger when the transcriber and encoder are two different people.

The challenges posed by bifurcating transcription and encoding come into sharper focus with projects like the CCL. As noted above, the success of the CCL project ultimately rests on the rapid expansion of a corpus of reliable transcriptions prepared to the highest standards of paleographic accuracy and properly encoded for use with the CCL search engine (designed to operate well across texts in Latin and with irregular orthography) and collation software. Since its inception, furthermore, the CCL team has learned that the burden of encoding transcriptions in an independent process is great, as it must be for many such large, complex projects. In most instances, encoding requires a continuous supply of paid labor and training and supervision of the labor force; even when encoders are available, the time required to produce a properly encoded text can be disproportionately long when compared to the time invested in transcription. Without some means of automating at least a portion of the encoding, many projects are going to be vulnerable to failure when there is insufficient funding for encoders. While we anticipate that there would be peculiarities in individual manuscripts that require considered manual intervention in the markup, it is also evident that by far the greatest portion of CCL encoding is predictable and formulaic.

What the CCL project needed was a tool that would allow the transcriber—the person with the greatest understanding of the manuscript and the particular text—to enter the appropriate markup from a set of clear options. Such a tool would ensure greater consistency and accuracy in both content and markup. Additionally, automating the structural markup would allow the guardians of the CCL to devote attention to the peculiar instances that require additional resolution of encoding problems. T-PEN answers that need for integration of the transcription and encoding processes. In the T-PEN environment, a transcriber—even one with no knowledge of xml or TEI, can supply the basic markup of the text using the TEI protocols explicated in the CCL's Guidelines for Encoding for the CCL.   Using that markup, the CCL team is able to rapidly complete the preparation of the digital transcription so that the file will validate against the CCL's encoding schema, will be displayed properly in the CCL search and collation software, and can be manipulated properly on the CCL website by registered account holders who wish to contribute a translation or annotation to be automatically linked to a particular textual segment.

7. T-PEN's XML Encoding feature

A project leader or individual transcriber can create a set of markup elements in T-PEN for insertion into the transcription as the transcriber (or other collaborator) works through the text. There are two general ways to initiate XML buttons. The first is to link an XML schema to the transcription project, after which T-PEN will create a button palette that can be manually refined: a button set generated directly from a schema will likely have header elements and other structural elements that a transcriber may not want to deploy in a transcription. In the case of the CCL, it proved advantageous to reduce the 63 markup elements validated in the schema to 23 that are most commonly used to identify textual and paleographical features.


Projects
Manuscripts
Collaboration
Opti
Link a schema to this project
Link XML
Validate XML for this project
Import Buttons from Linked Schema
Button Management

Project Options

Import XML schema, validate projects, and customize buttons.

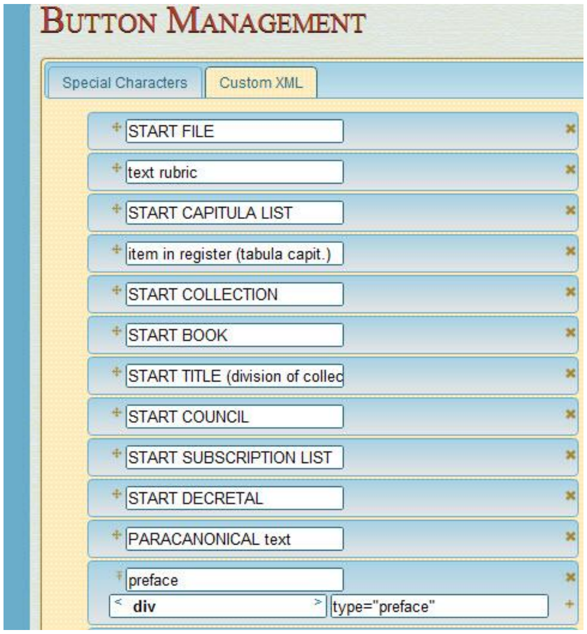
URL of schema:

enter complete file URL

The Special Character and Custom xml Tags you define will remain specific to each project. These buttons are accessible on the transcription pages. Characters assigned to the numbered buttons can be inserted simply by holding CTRL and pressing the corresponding number on the keyboard.

Figure 7 Detail of the Options tab in T-PEN's project management

The buttons can be customized to have simple, recognizable names (such as traditional scholarly terms). Buttons for XML-markup can also be created manually. In both cases, T-PEN allows the button designer to distinguish between a button description and the XML tag element name. This distinction allows user to identify quickly the function of the XML button rather than remembering what the TEI configuration may reference. For example, most transcribers will recognize that the button “gloss” would be used to identify textual elements in the manuscript that are commentary upon or clarification of the original text. They are not likely to be familiar with the CCL encoding for such an element: <note type=”gloss” target=”” place=””>. The XML buttons are named in the Transcription UI as the description (in this case, “gloss”) although the full tag syntax is inserted into the text. The T-PEN button management feature supports specification for each XML element to the attribute level. Once the buttons for a project are set, they are available through the "XML Tags" button in the transcription UI. By clicking on the XML Tags button, a transcriber can reveal the encoding buttons that have been created for the project; the buttons are shown immediately below the transcription box.



START FILE

text rubric

START CAPITULA LIST

item in register (tabula capit.)

START COLLECTION

START BOOK

START TITLE (division of collec

START COUNCIL

START SUBSCRIPTION LIST

START DECRETAL

PARACANONICAL text

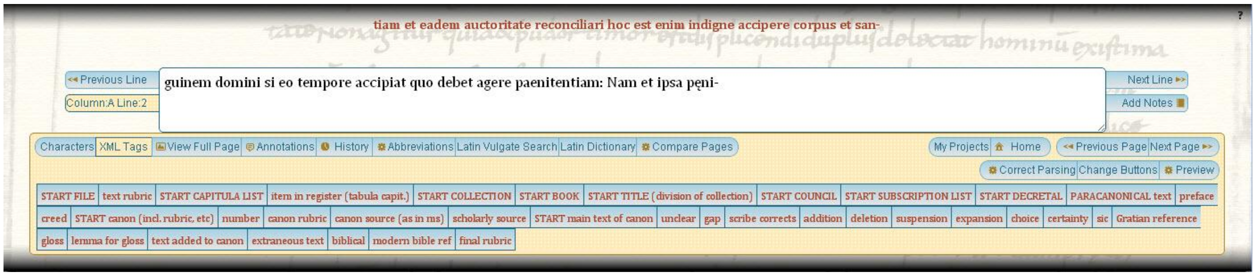
preface

< div > type="preface"

Figure 8 Detail of the Button Interface, where individual buttons can be created, modified or deleted

A third option for creating an XML button palette is to copy them from another project. This can be particularly useful if a collection of transcription projects are following the same encoding protocols and may become part of a larger database. This is certainly the case for CCL projects. Since all transcriptions completed in T-PEN for a CCL project will eventually be incorporated into the CCL database, sharing an XML button palette that contains the CCL XML schema ensures a high level of quality control in the application of correctly formed and consistently entered markup.

In this case (Figure 9), the transcription project, **Bern, Burgerbibliothek, MS 425**, is a CCL designated project and so has adopted the CCL's XML encoding schema.

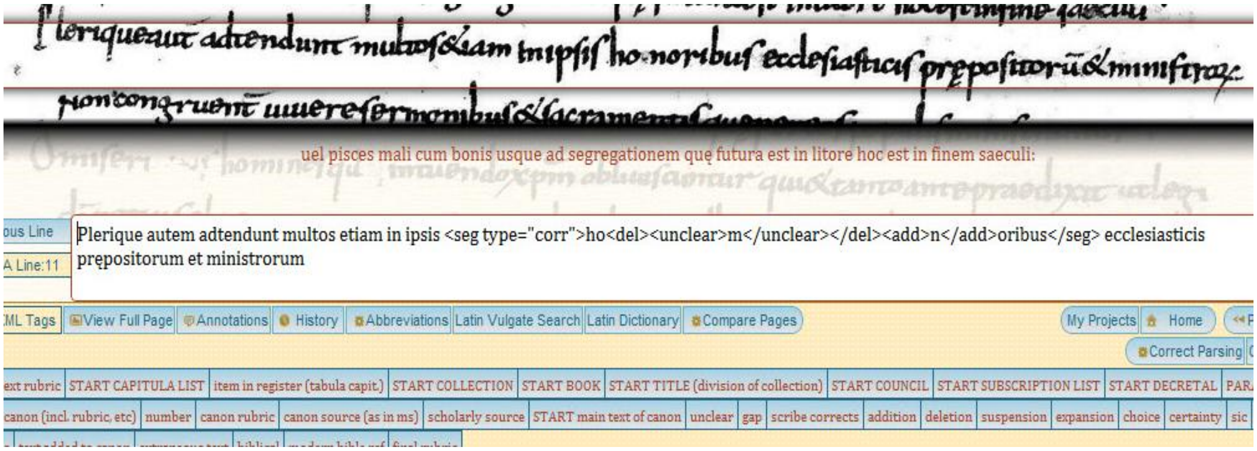


tiam et eadem auctoritate reconciliari hoc est enim indigne accipere corpus et san-

guinem domini si eo tempore accipiat quo debet agere paenitentiam: Nam et ipsa pęni-

Figure 9 : T-PEN's Transcription UI with the XML buttons revealed

Once an XML button palette is in place, the transcriber can begin to use it to insert XML tags into the transcription data stream. The transcriber (or other collaborator) simply clicks on the button and the XML tag is inserted wherever the cursor is positioned.

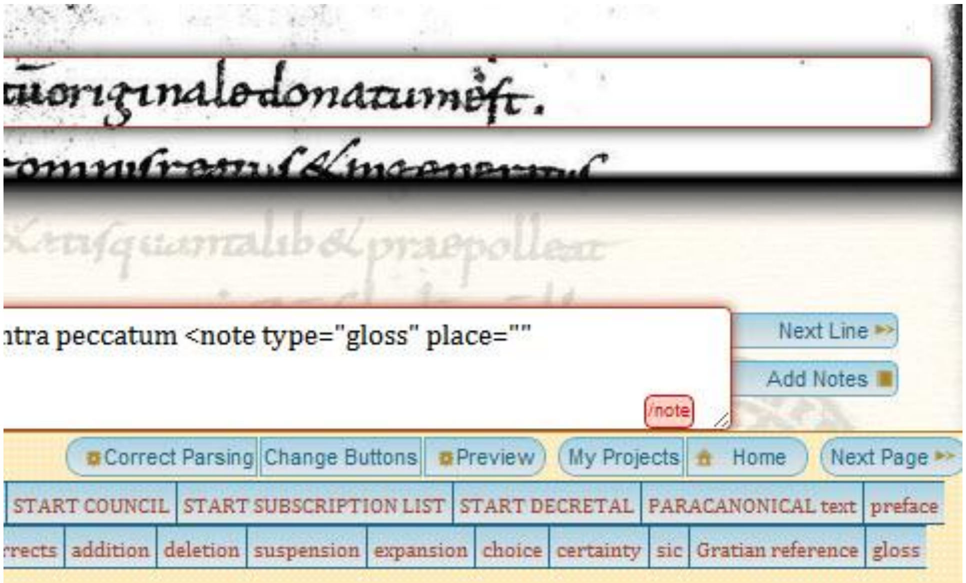


uel pisces mali cum bonis usque ad segregationem quę futura est in litore hoc est in finem saeculi:

Plerique autem adtendunt multos etiam in ipsis <seg type="corr">ho<del><unclear>m</unclear></del><add>n</add>oribus</seg> ecclesiasticis prępositorum et ministrorum

Figure 10 A transcribed line with XML tags inserted in the text

Clicking on an XML button inserts the open tag of the element. When that happens, T-PEN responds by displaying a closing tag (in red) for each opened element at the bottom right of the transcription text box. To insert the closing tag, the user simply clicks on that relevant tag. The "closing" reminder disappears. Closing tags persist from line to line and from page to page, until the user selects the tag or decides to delete the reminder.



ıtra peccatum <note type="gloss" place=""

Figure 11 Example of a closing tag reminder

An encoded transcription can be exported as an XML file and immediately imported into any XML editor. The export options also provide some basic transformation of XML elements if the user wishes to export an encoded transcription as a PDF, RTF or HTML file. It should be emphasized that T-PEN does not actually demand any encoding at all. This may sound like a heretical position to take in the world of digital humanities, but T-PEN expects a wide variety of users from those who encode everything to those who simply want to generate an electronic transcription that can be imported into word processing documents or HTML pages.

8. Completing "auto-encoding": T-PEN's Switchboard feature

As observed earlier, T-PEN connects transcribers and the resources of digital repositories. In a more substantial way, transcription data created in T-PEN will gain its full scholarly meaning either through refinement in other applications (such as an electronic typesetter for editions), or when it is aggregated with other similar texts. For the CCL, T-PEN provides a means to create accurate transcriptions, which will become part of the larger CCL textbase on the CCL's website. This aggregation (or integration) poses a procedural challenge. Good interoperability not only ensures data is structured in a way so that it can be shared, but that sharing can be done in a simple and iterative manner. Because T-PEN encourages large, collaborative projects that likely employ XML schemata, it made sense to offer a further set of "connections" to external, digital projects by making it possible to export transcriptions directly into their data stores.

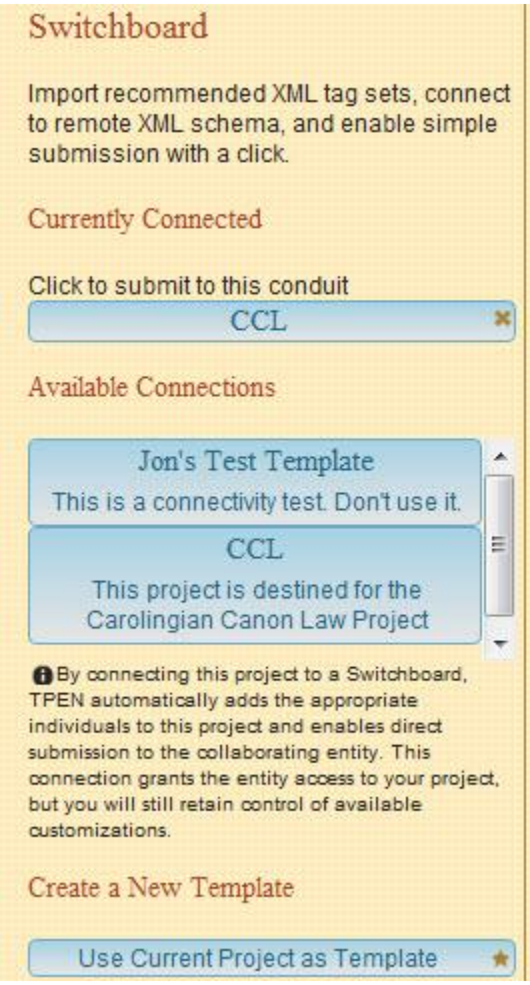
We therefore developed a feature called "Switchboard." On the Switchboard page, a project leader may designate a T-PEN project as destined for a specific, external project, by clicking a button that has been prepared in co-ordination with the T-PEN developers. Once that button is clicked, Switchboard connects the T-PEN transcription project and the external project by doing three things:

(1) Designated users from the large external project are automatically added as project collaborators in the T-PEN project, to facilitate whatever processing might be needed to transform the transcription into a file suitable for the external project. The original leader of the T-PEN transcription project retains that status, but the users added by Switchboard can monitor the progress of the project, and even modify its features or the transcription itself.

(2) T-PEN automatically associates the button palette that instantiates the external project's XML schema with the designated T-PEN transcription project.

(3) The T-PEN transcription project may "send" the transcription through Switchboard to the external project. The actual process is that T-PEN either alerts the external project's managers by email that a transcription is ready for extraction from T-PEN and import into its datastore, or responds to a web call provided by the external project. T-PEN creates an URL

that configures the transcription to be exported according to the protocols of the project using T-PEN's export feature. That URL is embedded in the alert email or web service interaction, which allows the external project to download the transcription or import it directly into their datastore..



Switchboard

Import recommended XML tag sets, connect to remote XML schema, and enable simple submission with a click.

Currently Connected

Click to submit to this conduit

Available Connections

Jon's Test Template

This is a connectivity test. Don't use it.

CCL

This project is destined for the Carolingian Canon Law Project

By connecting this project to a Switchboard, TPEN automatically adds the appropriate individuals to this project and enables direct submission to the collaborating entity. This connection grants the entity access to your project, but you will still retain control of available customizations.

Create a New Template

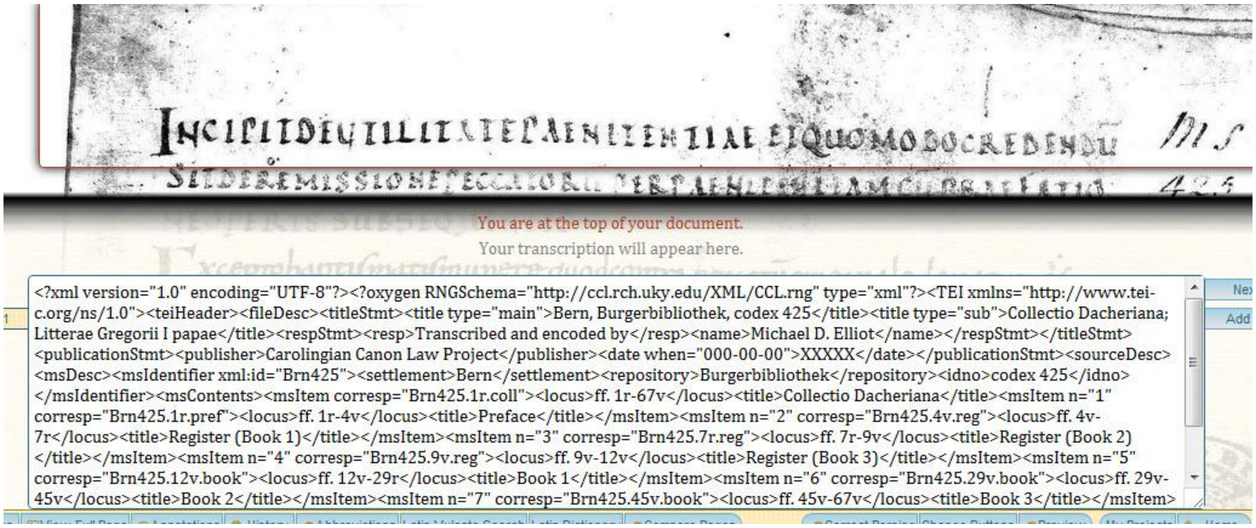
Figure 12 T-PEN's Switchboard UI in Project Management

T-PEN's Switchboard acts as a quality control mechanism in the same way that its collaborative environment does: it allows interested parties to verify and correct transcriptions and xml markup. It also reduces the steps needed to transfer the transcription data into another web environment or application. A user does not have to remember how to configure the export feature in ways specific to the external project’s needs. That has all been configured in advance by the external project managers instead.

In conjunction with the “auto-encoding” feature of T-PEN (the XML button palette), Switchboard also has the potential to increase greatly the sustainability of digital projects. While the term “sustainability” usually is used in the context of data preservation and curation, we have observed that, in the present stage of digital humanities evolution, it tends not to be data, but projects, that wither on the vine. The financing and personnel required to sustain ambitious digital projects are in short supply, and there are too many sad tales of impressive launches followed by lamentable stasis or, worse, disappearance. One of the results of the partnership between the CCL and T-PEN is that we were bringing together an established digital project facing the questions about long-term growth, steady expansion of its database, and feasible access to financial and labour resources. Once Switchboard was built, the CCL was able to produce programming that completes with minimal human intervention the encoding of a transcription. The resulting file can then be published on the CCL with a simple paste operation and one button click; it is then available for the full range of CCL software operations. The CCL post-export processing script may serve as a model for other projects with similar needs. The problem that we had to solve was how, on the one hand, to keep the XML-button palette in T-PEN to a manageable size, and to keep them in strict correspondence to the textual and paleographical features an xml-innocent scholar would know how to recognize and mark, and, on the other hand, to supply the plethora of markup required for files to validate against a sophisticated TEI schema.

9. Transcription encoding versus structural encoding

As all users of TEI know, there is much more to XML encoding than simply marking up portions of text with the appropriate tag. XML not only allows one to describe the content of a text, but to describe the text's structure as well. XML also allows intertextual and even inter-resource references to be embedded into the encoded text. This type of structural encoding is often essential, but can sometimes to be tangential to the task of transcription. Consider the basic question of a TEI header: while crucial for encoding the relevant metadata of the transcription project and required for file validation, it can easily become unwieldy when incorporated into a transcription, as figure 13 clearly demonstrates:



You are at the top of your document.
Your transcription will appear here.

```
<?xml version="1.0" encoding="UTF-8"?><?oxygen RNGSchema="http://ccl.rch.uky.edu/XML/CCL.rng" type="xml"?><TEI xmlns="http://www.tei-
c.org/ns/1.0"><teiHeader><fileDesc><titleStmt><title type="main">Bern, Burgerbibliothek, codex 425</title><title type="sub">Collectio Dacheriana;
Litterae Gregorii I papae</title><respStmt><resp>Transcribed and encoded by</resp><name>Michael D. Elliot</name></respStmt></titleStmt>
<publicationStmt><publisher>Carolingian Canon Law Project</publisher><date when="000-00-00">XXXXX</date></publicationStmt><sourceDesc>
<msDesc><msIdentifier xml:id="Brn425"><settlement>Bern</settlement><repository>Burgerbibliothek</repository><idno>codex 425</idno>
</msIdentifier><msContents><msItem corresp="Brn425.1r.coll"><locus>ff. 1r-67v</locus><title>Collectio Dacheriana</title><msItem n="1"
corresp="Brn425.1r.pref"><locus>ff. 1r-4v</locus><title>Preface</title></msItem><msItem n="2" corresp="Brn425.4v.reg"><locus>ff. 4v-
7r</locus><title>Register (Book 1)</title></msItem><msItem n="3" corresp="Brn425.7r.reg"><locus>ff. 7r-9v</locus><title>Register (Book 2)
</title></msItem><msItem n="4" corresp="Brn425.9v.reg"><locus>ff. 9v-12v</locus><title>Register (Book 3)</title></msItem><msItem n="5"
corresp="Brn425.12v.book"><locus>ff. 12v-29r</locus><title>Book 1</title></msItem><msItem n="6" corresp="Brn425.29v.book"><locus>ff. 29v-
45v</locus><title>Book 2</title></msItem><msItem n="7" corresp="Brn425.45v.book"><locus>ff. 45v-67v</locus><title>Book 3</title></msItem>
```

Figure 13 The Bern Project with a TEI header inserted in first line of the transcription

This looks unwieldy and, indeed, from a transcription perspective, it is. This issue demonstrates the limits of encoding in T-PEN: it has not been designed to act as an XML editor, but simply to make XML encoding possible where that task is directly related to the task of transcription.[1] The result is that T-PEN will always produce an incomplete XML file. To address this limit and to assist the encoding process as much as it can, T-PEN supports the uploading of a TEI header. The header can be added to the project (and subsequent uploads write over any existing header), but it cannot be edited in T-PEN. At export, users can select to include the header in the output file.

[1] The T-PEN development team has been exploring the possibility of interoperating with a web-based XML editor, in which encoded transcriptions could be easily sent to the editor via a web service and imported after editing. The most likely candidate for this task is the ANGLES project, proposed by UMITH. That project has yet to move beyond the prototype phase, but when it does a future version of T-PEN could certainly make use of it.

Even when the problem of the header and its content is addressed, there remains a substantial quantity of markup that will not, and should not, be undertaken by most transcribers. For example, TEI <div> tags often require <p> tags, and <text> tags may require encompassing <group> tags, to complete a valid hierarchy in the structure of the markup. These tag elements can be added to the XML file once it is exported from T-PEN. In the case of the CCL, our post-export script adds header information that is construed from the metadata available in the T-PEN project, and also adds the predictable structural encoding as is appropriate, by determining which textual and paleographic markup is present in the file and what empty tags are needed to complete it. Because the CCL identifies each textual unit with a unique, alphanumeric, xml identifier, that is constructed from the siglum assigned to a particular manuscript, the folio on which the text appears, and its position on the folio, we also include in the script a process that generates the needed xml:id formulae and places them appropriately in the file. The script matches the T-PEN identification of the manuscript with the CCL's list of manuscripts to find the siglum associated with the manuscript, finds the T-PEN record of individual image names and transforms them into TEI folio notation, and counts the number of textual units on the manuscript page to determine their numbered positions. In short, the CCL is now able to handle most (it will never be all) transcriptions generated in T-PEN and transferred to the CCL with Switchboard with a largely automated process for inserting almost all the necessary TEI 5 markup. This is the single greatest advance in sustainability of the CCL.

10. Observations: What we have learned

After a full year of development, the T-PEN development team has come a long way in its thinking. We have advanced in our thinking about the general features of web-based software. Further, we have also seen our views change on what features constitute an effective transcription tool that can be used by scholars of all different interests and persuasions. While the two principal investigators are both medievalists by training, and the initial use cases have focused almost exclusively on pre-modern European manuscripts, our aim has to been to create a tool that can meet the needs of modern historians, literature specialists of any era, and even those who work with unpublished documents related to government and public policy. In terms of lessons learned, there are two specific areas.

(a) Public versus private repositories or, what to do about local uploads

Part of the mandate for the NEH grant was to create a mechanism for users to upload their own images. This was considered an optional feature, as the primary focus of T-PEN has been to act as a conduit between digital repositories and transcribers. The main reason for pursuing a "repository" strategy was that it ensured use of any image would conform to the conditions of use established by the repository. In fact, when a repository agrees to give T-PEN access to its resources the conditions of use statement is incorporated into T-PEN, and when a user gains access to manuscript from that repository for the first time, they must agree to those conditions. Related to this issue is the fact that T-PEN does not store or make copies of any image that a transcriber makes part of any project. When a user requests an image, it is fetched from repository itself. T-PEN does use a temporary cache system where an image is retained in order to speed up page loads, but there is no long term, persistent storage of images from repositories. T-PEN must not facilitate violation of or disregard for the Intellectual Property Rights associated with the digital images of any repository.

It thus became difficult to implement a general method for users to upload their own digital images of manuscripts. The practical reason was that the T-PEN server was hosting a mirror copy of the Parker on the Web collection, while we waited for the collection to be properly accessioned on Stanford University's SharedCanvas platform. This reduced the available storage space for individual uploads, a situation that has been since resolved after the end date of the NEH grant period. More substantially, however, was the problem of maintaining a commitment to IPR standards. We could not, in the first instance, identify a process which did not put T-PEN into a situation where we might be liable for illegal behavior of a T-PEN user. We consider individual copies of manuscript images to be used for private, scholarly purposes to fall under the "fair use clause" (US Code, Title I, sec. 107), but the question we needed to answer was: did making such a page image available on T-PEN contradict fair use?

Our answer was to add a restricted access condition in the manuscript object. When a user requests to upload a "private" collection of digital images of a manuscript, T-PEN's administrator designates it as a restricted manuscript. This means that the manuscript will appear on T-PEN's catalog of available manuscripts, but only assigned individuals can gain access to the images. All manuscripts digitized with NEH funds for the CCL are hosted by T-PEN and are treated as restricted manuscripts.

T-PEN has also deployed this restricted modifier to collections that may require a subscription. Since T-PEN cannot authenticate whether a user has subscription rights to a collection, T-PEN assigns a manuscript administrator to that collection, and users can then contact them (within T-PEN) to gain access. Once that administrator has determined the status of the user, he can provide access to one or more (or the whole collection) manuscripts. Currently, the request to upload a private digital manuscript is a manual operation. T-PEN, by the time of its final release version in April 2012, will have an automated process in place. This will require the T-PEN administrator to verify the contents (so as to ensure there is no inappropriate or illegal materials are being uploaded) before the manuscript is added to the manuscript catalog.

(b) Instantiations of T-PEN or, how can we help?

A second issue that has generated discussion amongst the development team is how to make T-PEN available. T-PEN is currently a web-based, freely available application and the CDT and the CCL are committed to keeping that way. Our initial vision was that anyone could install their own copy of T-PEN on their server. The condition was that all transcriptions generated on that site would have be aggregated with the data store at the CDT, so that we could provide a comprehensive search engine. T-PEN can be installed on any server. When the codebase will be published as open source on *Sourceforge (in April 2012)*, there will be a configuration file that will make the process relatively simple (provided that the server has both Apache and Tomcat running). We do not expect a significant number of providers to take up this option. In most cases, T-PEN's own instance will be enough to provide good service to many repositories and large collaborative projects. In the initial proposal to the NEH, we recommended that the CCL would install its own version of T-PEN. By the end of the grant, we concluded that this was too onerous on the CCL since it increased their amount of maintenance without any clear benefits being gained. With Switchboard activated, the CCL has gained a web application that has been configured for its specific needs. CCL did not have to reconfigure its data models, change its own workflow in order to exploit the services of T-PEN. In the end, it proved beneficial for the CCL to continue to use T-PEN hosted at the CDT. This flexibility in configuration of output, we surmise, will be very useful to other large, collaborative projects.

Conclusion

This paper has demonstrated that T-PEN as a digital tool not only fully addresses the need to integrate scholalry transcription and encoding, but it offers a methodology that can be applied to research projects ranging from a transcriber working in isolation on a critical edition to a large collaborative projects where there may be hundreds of transcribers all working from a common encoding schema. We have been able to accomplish this because we have always had a model of scholarly practice in mind that considers even the solitary transcriber to be part of a larger community of scholars. Hence, be it one or a thousand transcribers at work, they all need to accomplish the same task. T-PEN assists in that task to ensure scholalry excellence and accuracy, but it also permits sharing data and supports ways for those outputs to be aggregrated at, or integrated into, a larger textbase. All transcribers need access to the document they wish to transcribe. T-PEN creates a digital workspace that takes full advantage of what digital repositories have to offer. All transcribers need to annotate or encode. T-PEN integrates transcription with encoding so that they work in tandem and ensure no loss of detail in the process. All transcribers need tools to assist them in their work. T-PEN provides some key resources to asssist in the transcription and more importantly has a framework to add other tools when the need arises. All transcribers need to have their work available for other contexts, be it part of an article or monograph, a published or digital edition, or a large textbase of related texts. T-PEN makes getting transcriptions of out its data store an easy task and has gone to significant lengths to ensure that export is appropriately configured for large projects like the CCL.   There is clearly more development opportunities for the future but thanks to the funding of the NEH, T-PEN can become a usable tool to asssist in the scholarly reading and analysis of unpublished documents.

References

Burghart , Marjorie; Rehbein, Malte. 2012. "The Present and Future of the TEI Community for Manuscript Encoding." Journal of the Text Encoding Initiative 2. URL: http://jtei.revues.org/372.

Driscoll, M.J. 2006. "P5-MS: A General Purpose Tagset for Manuscript Description." Digital Medievalist 1. URL: http://www.digitalmedievalist.org/journal/2.1/driscoll/

Firey, Abigail. 2010. "The Carolingian Canon Law Project: A Collaborative Initiative." NEH-ODH White Paper. URL: http://www.neh.gov/ODH/ResourceLibrary/LibraryofFundedProjects/tabid/111/Default.aspx (keyword: Carolingian).

Ginther, James R.; O'Sullivan, Tomas. 2010. *The Electronic Norman Anonymous: An Integrated Edition of Test and Manuscript Page Images*. St Louis: Center for Digital Theology. URL: www.normananonymous.org

Haslhofer, B; Simon, R; Sanderson, R; Van de Sompel, H. 2011. "The Open Annotation Collaboration (OAC) Model." *Proceedings of the Workshop on Multimedia on the Web*. Graz: i-Media. URL: http://arxiv.org/pdf/1106.5178.pdf

Likforman-Sulem, Laurence; Zahour, Abderrazak; Taconet, Bruno. 2007. "Text Line Segmentation of Historical Documents: a Survey." *International Journal on Document Analysis and Recognition* 9, 2-4:123-138.

Sanderson, Robert; Albritton, Benjamin; Schwemmer, Rafael; Van de Sompel, Herbert. 2012. "SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination." *International Journal of Digital Libraries* 13, forthcoming. Pre-print: http://arxiv.org/ftp/arxiv/papers/1110/1110.3687.pdf